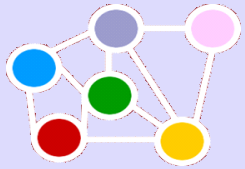
A la recherche du document perdu

on the de-construction of the 'document' notion

in emerging digital library settings

and on the vital importance of open access and open sources

Prof. Dr. Stefan Gradmann
Humboldt-University Berlin / Europeana
stefan.gradmann@ibi.hu-berlin.de
http://www.ibi.hu-berlin.de/institut/mitarbA-Z/professoren/gradmann



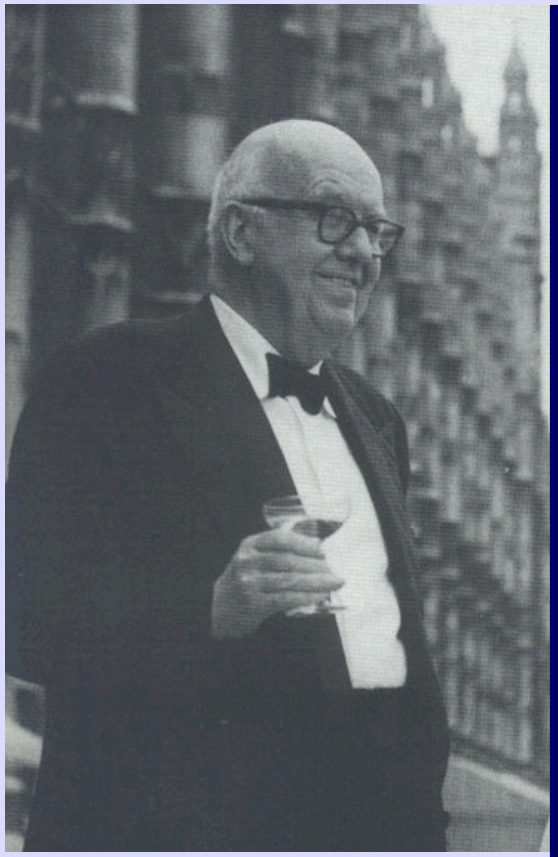
Overview

- **Congratulations, background & apologies**
- The **'two cultures'** distinction
- Context: an **evolving information continuum: beyond emulation mode**
- **Specific consequences** for **digital scholarly publication**, **open document models** and **standards**?
- What is **e-scholarship** and how does it specifically relate to its **object corpora / sources**?
- Consequences resulting from this specific difference regarding **'open source'** approaches?
- Some concluding words on the **re-constitution of the 'document' notion**, **OS**, **OA** and a last look at the **'two cultures'**

Presentator's Context: Digital Humanities, Digital Libraries & Open Access

- Definitely **not** a physicist!
- Background in **literary scholarship** (work on Joyce, Kafka, Arno Schmidt, Semiology, Greek Mythology), **Digital Libraries** and **Digital Semantics**
- Open Access Publishing (GAP, DINI)
- International advisor to "Our Cultural Commonwealth" (American Council of Learned Societies Report on Cyberinfrastructure for the Humanities and Social Sciences)
- Building the European Digital Library / **Europeana**
- Currently teaching Library and Information Science at Humboldt University / Berlin (Knowledge Management)
- And before starting I wish to seriously **congratulate this community** for long lasting as well as recent achievments!

The 'Two Cultures'



C.P. Snow in his Rede-Lecture of 07.05.1959 stated:

The breakdown of communication between the sciences and the humanities has led to the establishment of two distinct cultures of dealing with knowledge.

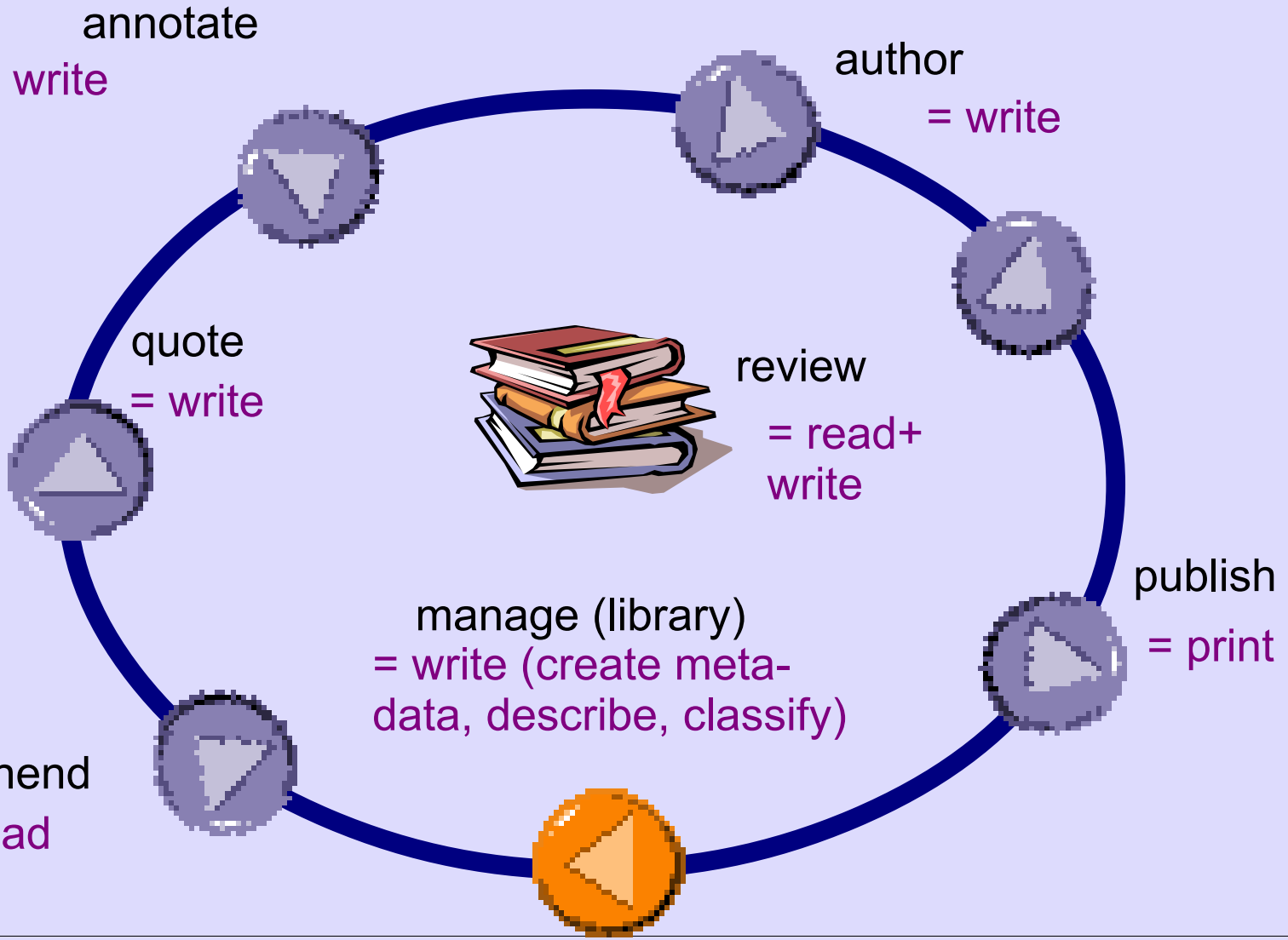
- 'Hard' Sciences - Empirical focus: finding routes towards a known target / intelligent retrieval strategies – "Explain"! "Measure"!

- Humanities - Hermeneutical focus: strive for 'knowledge' / finding 'reasons' – "Understand"!

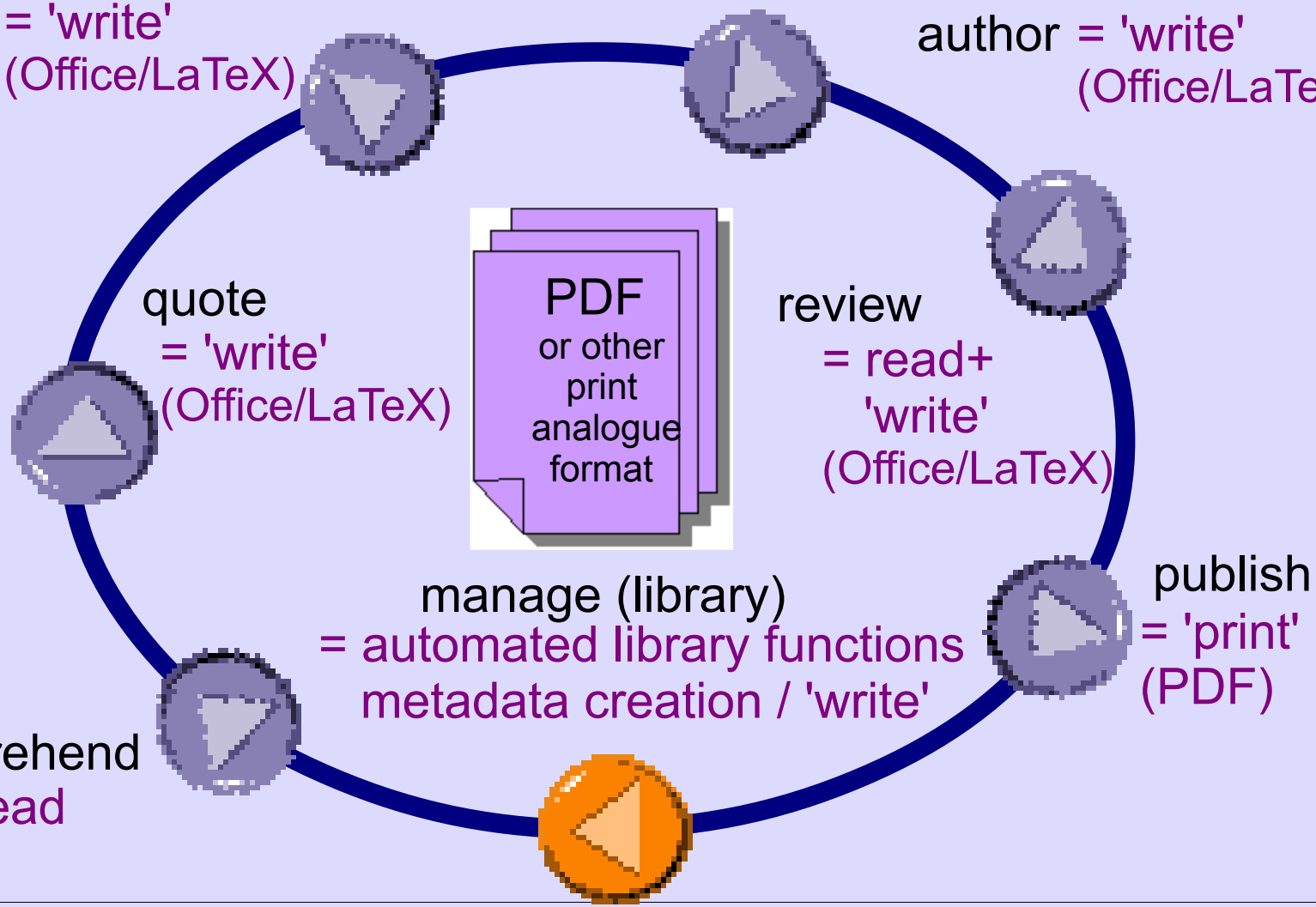
Understanding the differences: approaches to OA and 'documents'

- The division put in place by Snow is simplistic and inappropriate in many respects – it can be useful, however, to understand some of the specific aspects of the 'document' notion and of Open Access in the context of e-scholarship.
- There are two distinct cultures of publication and open access!
- Starting from W. McCarty: *„Academic publishing is one part of a system of highly interdependent components. Change one component [...] and system-wide effects follow. Hence if we want to be practical we have to consider how to deal with the whole system.“*

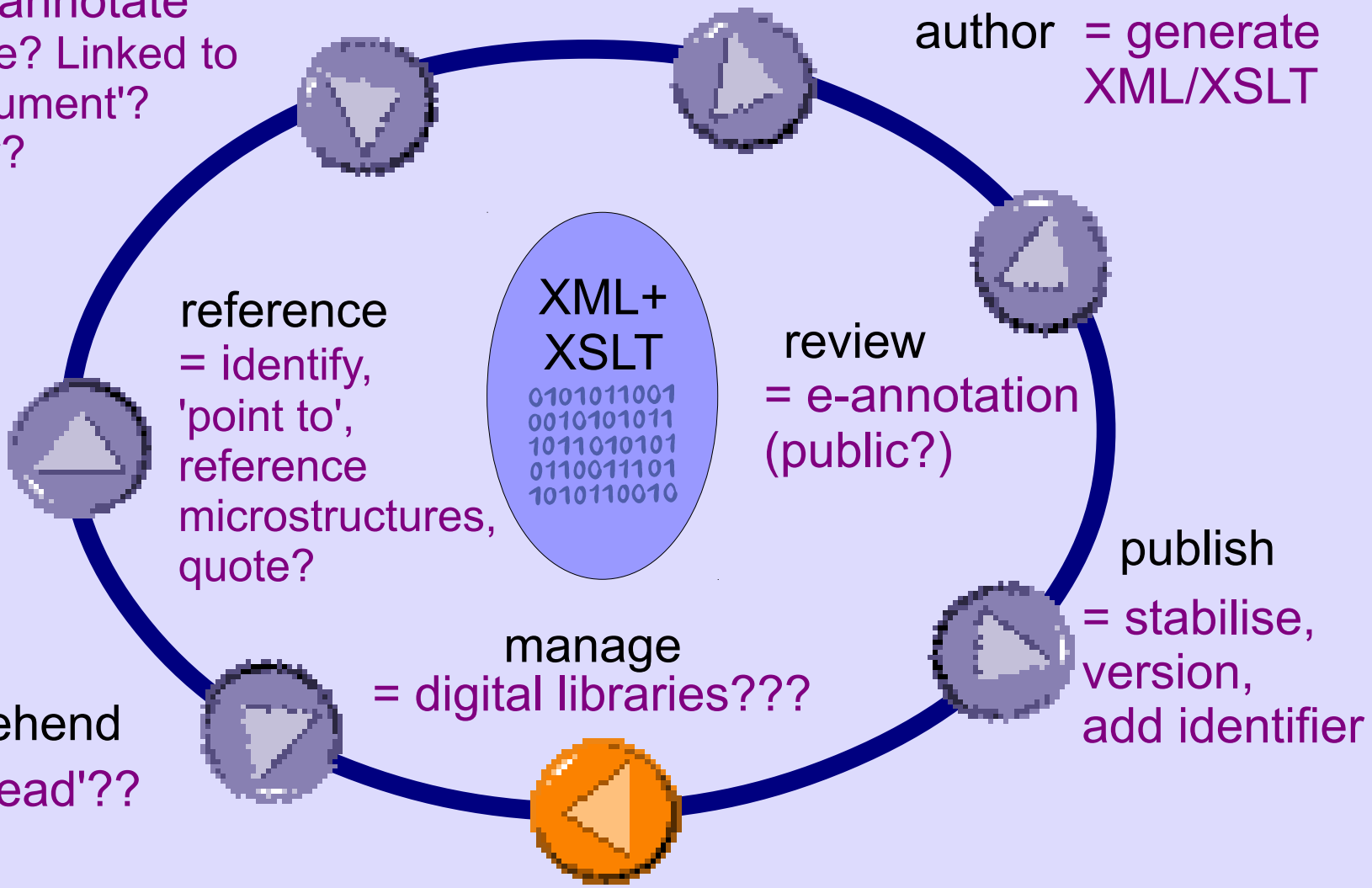
Linear Information Continuum using traditional cultural techniques



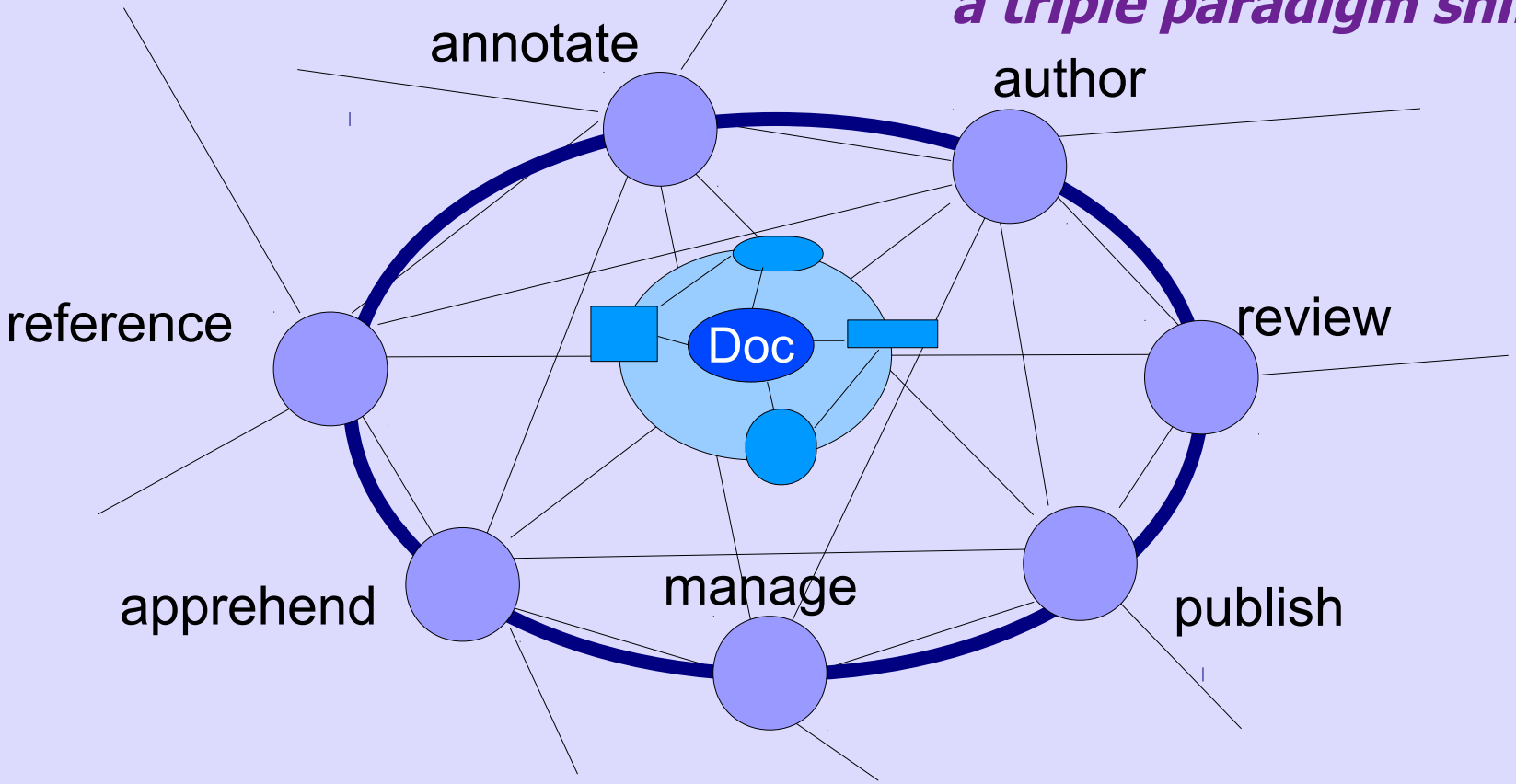
Linear Information Continuum Electrified
emulating traditional cultural techniques



Linear Information Continuum
... going digital



Scholarly Information Continuum
a triple paradigm shift



- Erosion of the linear / circular function paradigm
- Functionality is not any more entirely determined by traditional cultural techniques and related metaphors – and is not yet entirely determined by digital and still emerging technology
- De-Construction of the 'document' notion in a digital, networked context

Consequences of the triple paradigm shift for the humanities

- The Erosion of the linear / circular function paradigm only marginally affects the humanitues because of their 'monolithic' publication culture.
  - Journal publications as well as the related workflows and peer reviewing scenarios still play a less prominent role in our context.
  - Most authors in the humanities still basically work in isolated, autonomous settings, group authoring scenarios still tend to be exceptional.
- The decrease of functional determination by traditional cultural techniques does affect the humanities in many respects - none of these, however, being specific for the humanities.
- The De-Construction of the 'document' notion in digital, networked settings vitally affects the humanities in that it fundamentally changes the conditions of production and use of 'documents' and namely
  - Conditions of **apprehension** and **reuse**
  - Fundamental **signification modes of 'documents' seen as complex signs/sign clusters**

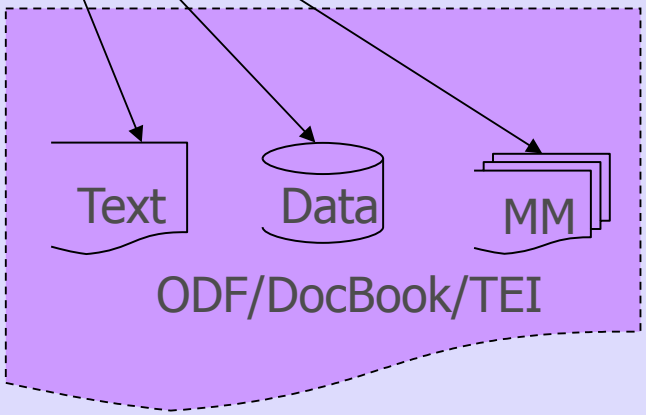
E-Science vs. E-Scholarship: different relations of research and 'documents'

- Signification and document modelling in OA related discussion up to now have basically been coined on the information model prevailing in the empirical sciences which in turn was based on the dissociation of resarch/data and publication:
  - **Research** => **'Results'** => 'Packaging' => Publication
  - Robust and not very complex 'container' modell
  - Electronic Science ≈ Electrified Science (e-science)
- Document modelling in the humanities and social sciences takes place in a substantially different information model:
  - **(Research <=> discoursive 'packaging')** => Publication
  - Resulting in complex document models heavily intertwined with core research operations
  - Complex signifier<->significate relations as constituents.
  - 'container' models are over-reductionist and inappropriate

Document perdu: quotations / references

„Author: Title.
p. 231“

„http://www.xyz.de/object#text.
chapter.paragraph.word.xml“
„http://www.xyz.de/object#MM.
image.<???>.xml“



- Is identification of networked electronic documents using constructs such as DOI/URN sufficient? How to 'point' to microstructures?
- Do object models such as MPEG or TEI provide adequate conceptual frameworks?
- And will we still replicate (quotation) or will we reference (pointer)

Open and standardised document models

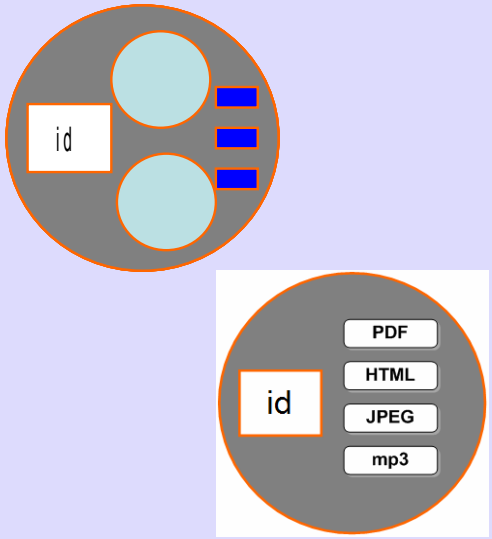
- Digital paradigm shift in the humanities vitally depends on open and non-proprietary techniques for document modelling and authoring
- This is even more evident if one considers not just isolated documents, but webs of interrelated documents pointing and referring to each other.
- This evidence is particularly striking if one considers the need to maintain coherent webs of documents over time for decades or even centuries
- Introducing document protection technology such as for DRM in such settings would create ridiculous and nightmarish functional scenarios!

Or – as one of your peers (Gigi Rolandi) put it at last year's PPA summit when asking himself how to analyse today's data in 2020:

- "The problem is NOT data and software preservation.
- The problem is the lack of a simple data model."

And this translates back to "document model" in the humanities

Compound Information Objects
As conceived in Object Reuse and Exchange (ORE)

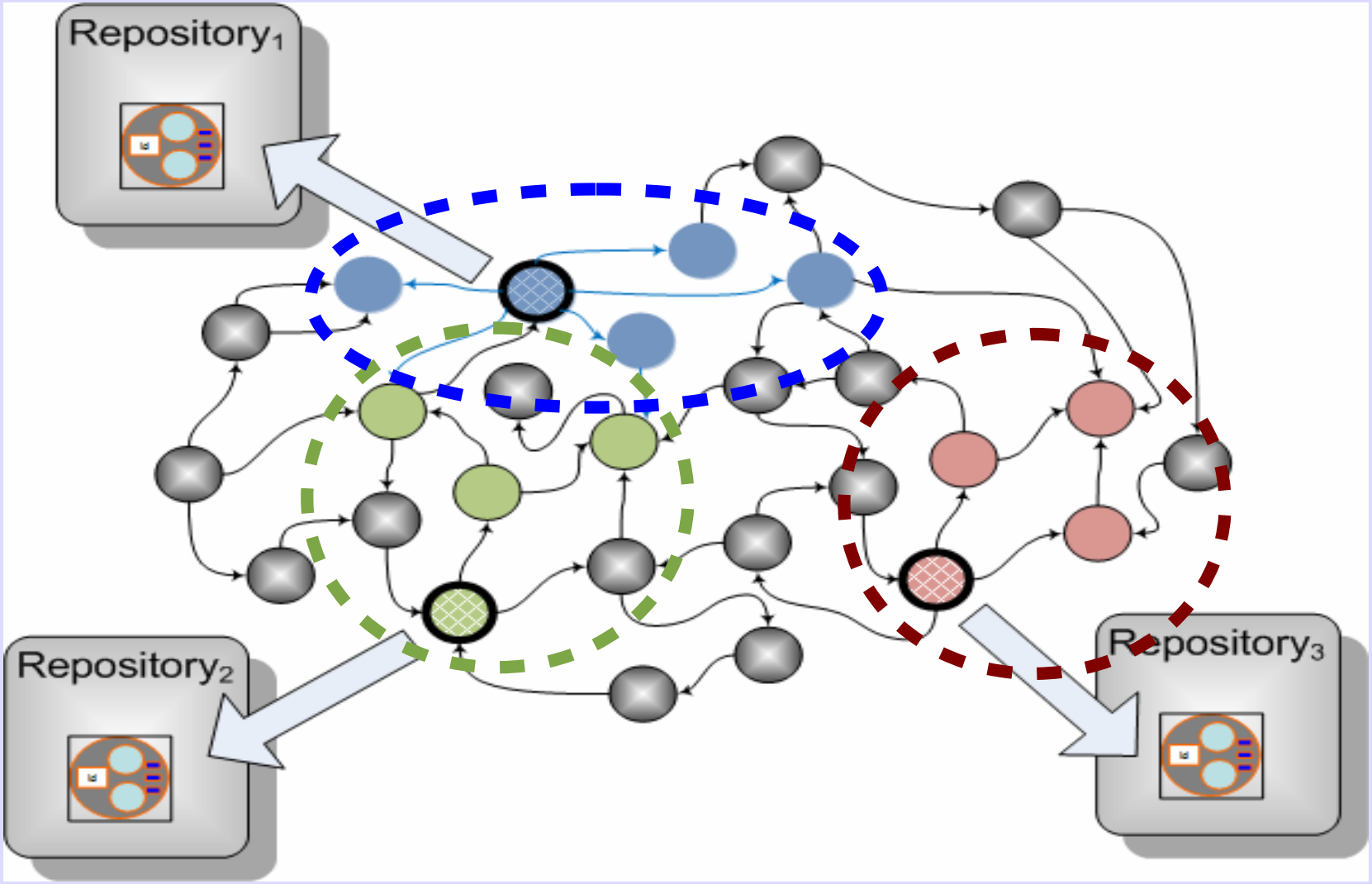


compound information objects

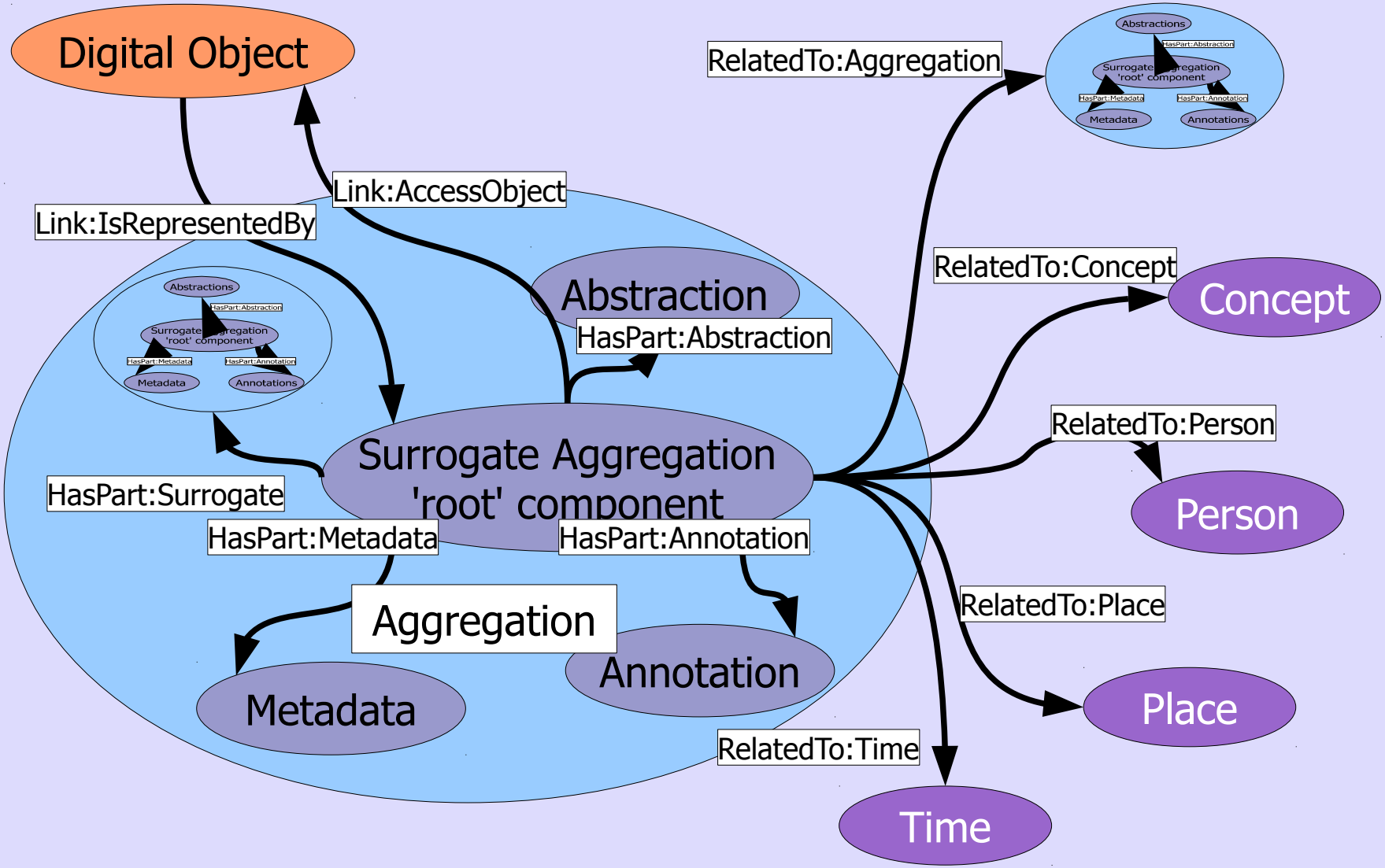
- Units of scholarly communication are compound information objects:
- **Identified, bounded aggregations of related information units that form a logical whole.**
- Components of compound object may vary according to:
  - Semantic type: book, article, moving image, dataset, …
  - Media type: PDF, HTML, JPEG, MP3, ...
  - Internal relationship: parts, views, …
  - External relationships

(Lagoze, Nelson, Van de Sompel 2007)

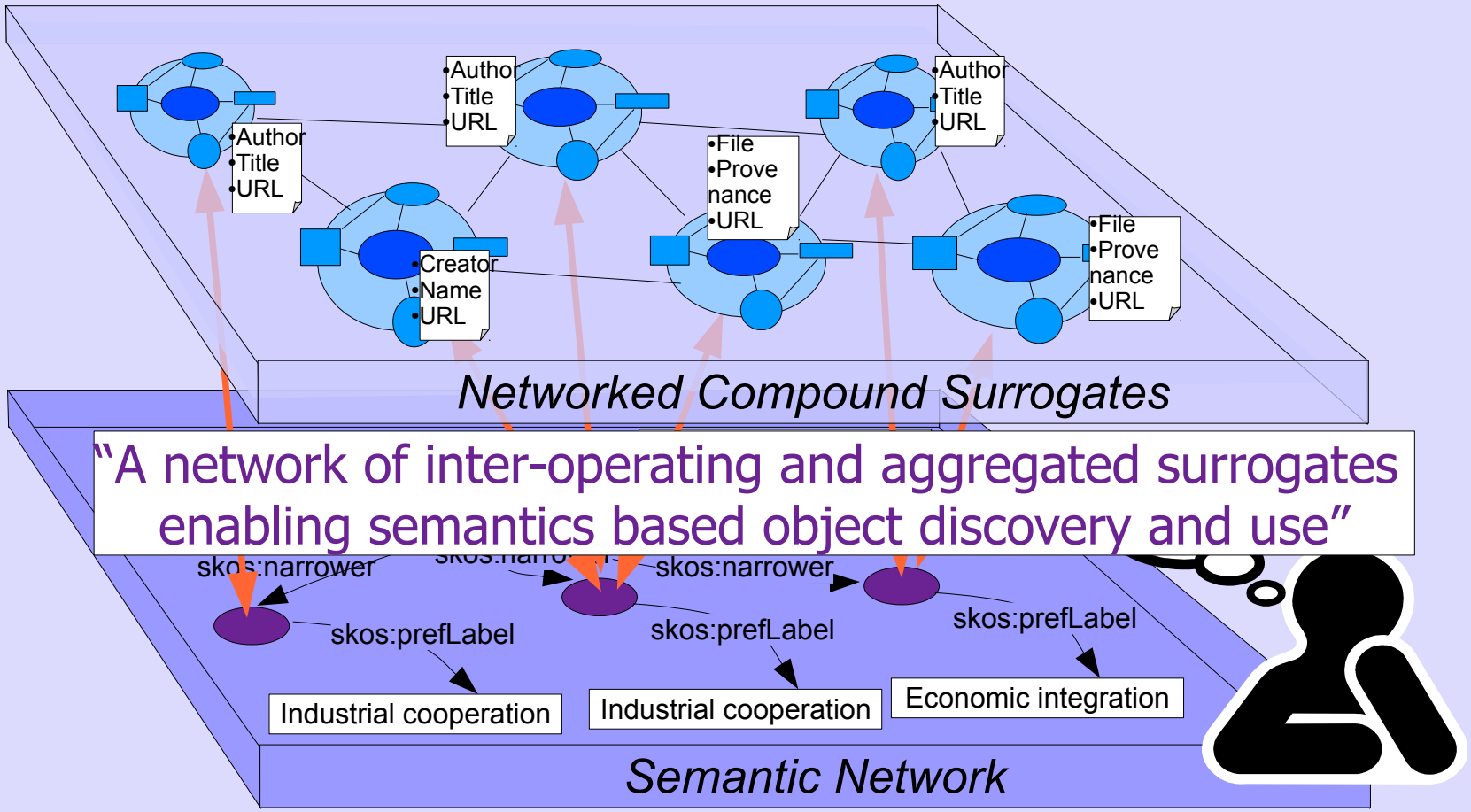
A similar model for web resource aggregations
Object Reuse and Exchange (ORE)

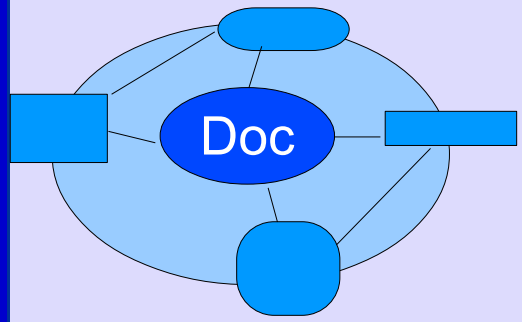


Object Model for Europeana Surrogate Aggregations



Document Objects, Metadata and Semantic Networks





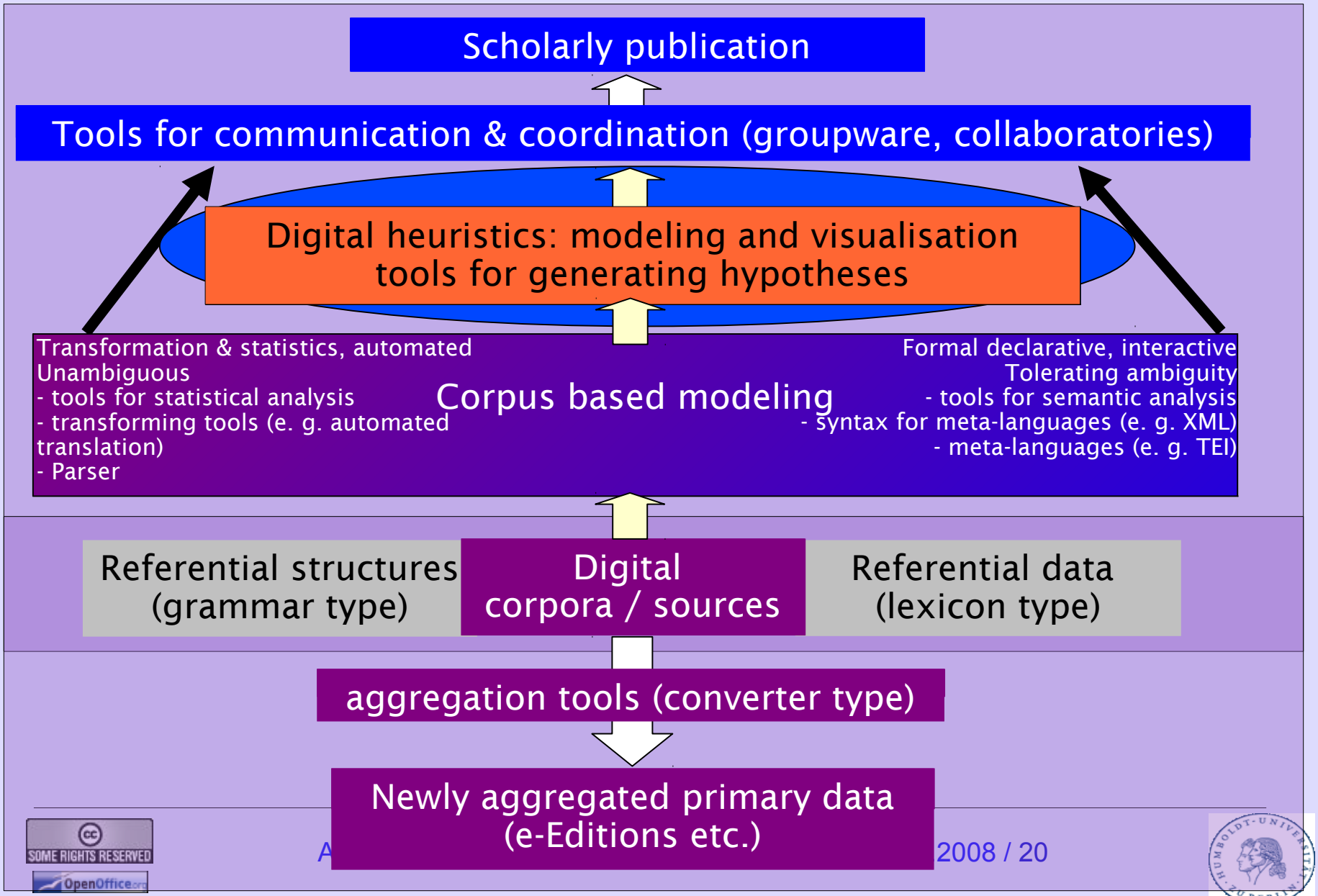
De-constructing the 'Document' Notion
The work of R.T. Pédauque

- RTP Doc (CNRS, http://rtp-doc.enssib.fr): Evolution of the 'document' notion in the passage from **printed** to **digital** to **web based** documents along three paradigms
  - **Form** (vu='Look at', morphosyntax), as material or non-material structured object
  - **Sign** (lu='read', semantics), as meaningful instance and thus both intentional and part of a sign system
  - **Medium** (su='Knowledge, Interpretation, Apprehension', Pragmatics) as a vector of communication, part of a social reality with constituting temporal and spatial processes of mediation
- RTP provides elements for re-constituting the document notion ...

A Second Look 'Documents' in the Humanities: ... after disintegration of the document notion

- ... but only **after** its current **disintegration**
  - caused by document resource distribution in networked settings
  - with no (or weak) cohesive forces to compensate the loss of constitutive linearity and integrity
  - that in turn was long guaranteed by closely coupled content and medium in the book culture!
- This de-construction process profoundly challenges humanities' scholars who are vitally rooted in a culture based on a traditional vision of 'documents'
- And I am even skipping the 'pandora box' of semiotics here to keep things relatively simple

Processing of source data in the Humanities: modeling and aggregation



An example of (proto)-DH and 'Hermeneutical Modeling' environment

EU-NSF-Projekt Cultural Heritage Languages Technology (CHLT)*

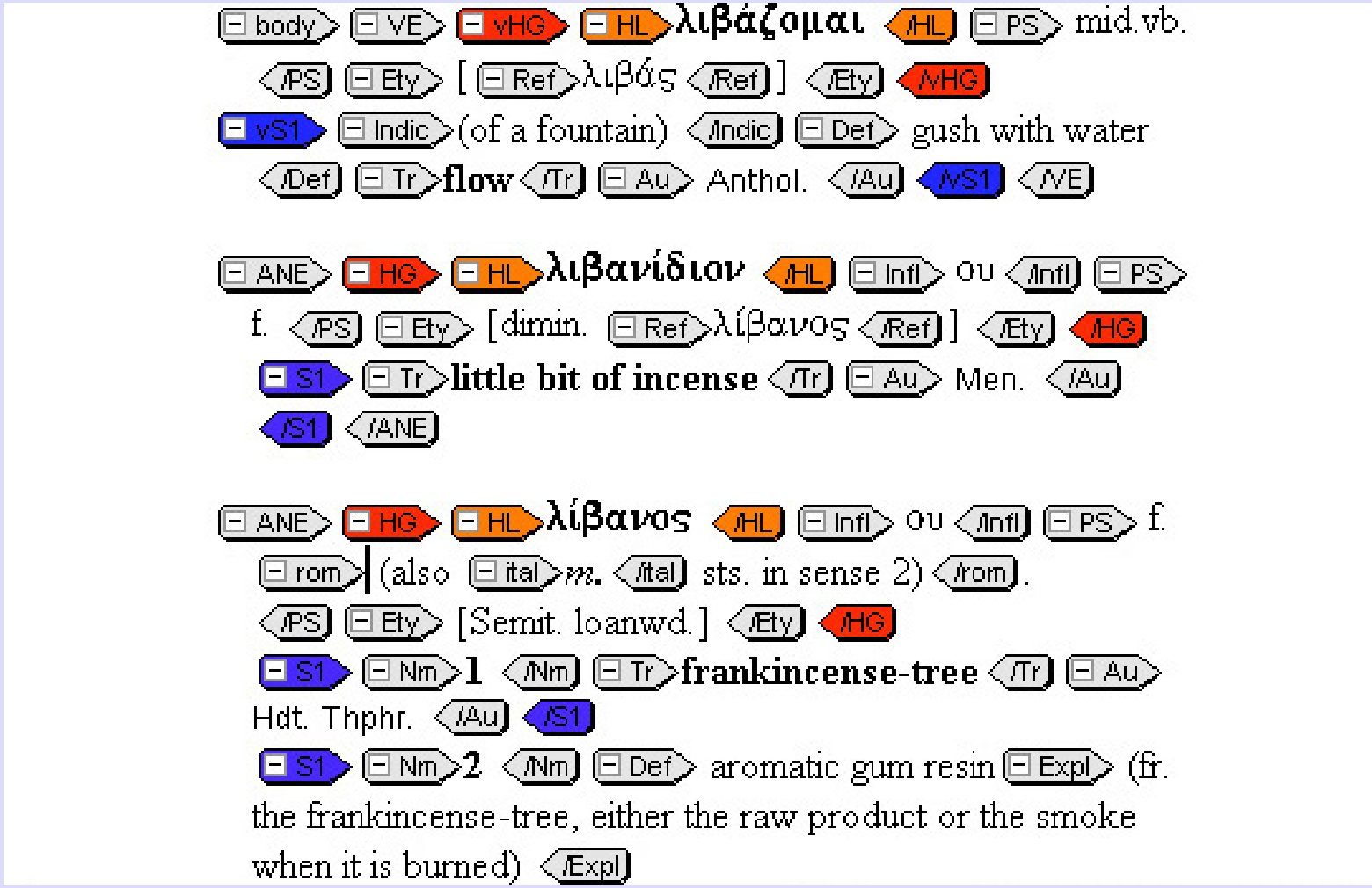
*Thanks to Bruce Fraser (Cambridge) and other participants in CHLT!

- Greek Lexicon Project (Cambridge): TGL => LSJ => TLG => GLP
- XML online-slips + XSLT
  - Supports annotation
  - Can be organised in multiple dimensions
  - Building block of a distributed, networked 'Collaboratory' (together with Perseus DL and others) for cultural heritage research work based on advanced technology
- Combined with tools for lexical clustering and visualisation of lexical distribution this is evolving into a fertile environment for generating text-related hypotheses.
- Illustrates well the methodological advantages of a clearly delimited and highly 'pre-aggregated' corpus.

Lexicon Code Fragment

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet href="../../../Lexicon/Dtd/lexicon.css" type="text/css"?>
<!DOCTYPE lexicon SYSTEM "file:///../../../../Lexicon/Dtd/lexicon.dtd">
<lexicon>
  <header>
       <file>   <title>Greek Lexicon Sample Page 1</title><editor>AAT</editor><date>7/4/04</date></file>
  </header>
<text>
       <body>
<ANE><HG><HL>λαγω<hyph/>βόλον</HL><VL><Lbl>also</Lbl><FmHL>λαγωβόλον</FmHL><Au>Anth.</A
      u></VL><Infl>ου</Infl><PS>n</PS><Ety><Ref>λαγώς</Ref>,
                        <Ref>βόλος</Ref></Ety></HG>
          <S1><Qualif>orig.</Qualif><Def>stick for throwing at hares<Expl>in
                        hunting; or simply as a mark of the
                        countryman</Expl></Def><S2><Tr>throwing-stick, stick</Tr><Au>Theoc.
                          Anth.</Au></S2>
          </S1></ANE>
         <ANE><HG><HL>λαγώδιον</HL><Infl>ου</Infl><PS>n</PS></HG>
          <S1><Def>young hare</Def><Tr>leveret</Tr><Au>Ar.</Au>
          </S1>
         </ANE>
         <ANE><HG><HL>λαγών</HL><Infl>όνος</Infl><PS>f</PS><Ety>reltd.
                        <Ref>λαγαρός</Ref></Ety></HG>
          <S1><Tr> flank, side, waist<Expl>of a person or animal, ref. to the
                        area betw. the ribs and the hip, or more generally, in sg. or pl., to the
                        middle of the body</Expl></Tr><Au>E. <NBPlus/></Au><S2><Tr>side <Expl>of a
                          mountain, a river</Expl></Tr><Au>Call. Anth.</Au></S2>
          </S1>
          <S1><Tr>recess, hollow<Expl>of a container, such as a cup, a
                        quiver</Expl></Tr><Au>Eub. Anth.</Au><S2><Indic>under an overhanging
                          rock</Indic><Au>Plu.</Au></S2>
          </S1><Ann><Para>how to translate Call.5.88 breasts and hips of Athena?
                or lagones more generally for the body, or the middle area of the body; perh.
                waist here. Sense 'womb', see Rev.Suppl., prob. doesn't exist. </Para><Para>at
                [...]
```

Lexicon Code Lemmatized



Lexicon Page Display

λαγω-βόλον, also **λαγωοβόλον** Anth. ου *n.* [λαγώς, βόλος] orig., stick for throwing at hares (in hunting; or simply as a mark of the countryman); **throwing-stick, stick** Theoc. Anth.

λαγώδιον ου *n.* young hare, **leveret** Ar.

λαγών όνος *f.* [reltd. λαγαρός] **1 flank, side, waist** (of a person or animal, ref. to the area betw. the ribs and the hip, or more generally, in sg. or pl., to the middle of the body) E. +; **side** (of a mountain, a river) Call. Anth. **2 recess, hollow** (of a container, such as a cup, a quiver) Eub. Anth.; (under an overhanging rock) Plu.

λαγῷος ον *adj.* [λαγώς] **of a hare** (ref. to the meat) Ar. || NEUT.PL.SB. (w. κρέα understood) hare-meat, cooked hare-meat dish Ar.

λαγώς λαγώ, also **λαγῶς** λαγῶ, ep. **λαγωός** οῦ Ion. **λαγός** οῦ *m. and f.* | acc.sg. λαγών (Ar.) | **1 hare** Hom. +; (as a type of timidity or cowardice, esp. in provbl.phrs.) Posidipp. D. +

λαθι-κηδής, Aeol. **λαθικᾱδής** ές *adj.* [λανθάνω, κῆδος] | acc.sg. λαθικάδεον | causing forgetfulness of care; **soothing** —*of a mother's breast* Il.; **banishing cares** or **pain** —*of wine* Alc. —*of Apollo* (*as healer*), *of medical knowledge* Anth.

λαθί-πονος ον *adj.* **forgetful of pain** or **trouble** S.

λαθι-πορφυρίς ίδος *f.* a kind of bird (app. which is hard to see, or is active only at night), perh., **shy purple-gallinule** Ibyc. | see also πορφυρίς

λαθί-φθογγος ον *adj.* causing forgetfulness of speech, **silencing voices** —*of death* Hes.*Sc.*

λαθιφροσύνη ης *f.* [reltd. φρονέω] (pl.) forgetfulness of common-sense, **madness** A.R.

λαθοίατο (aor.2 mid.optat.): see λανθάνω

λᾶθος εος *dial.n.* [reltd. λήθη] **forgetfulness, indifference** (*as the cure for love*) Theoc.

Annotation Support

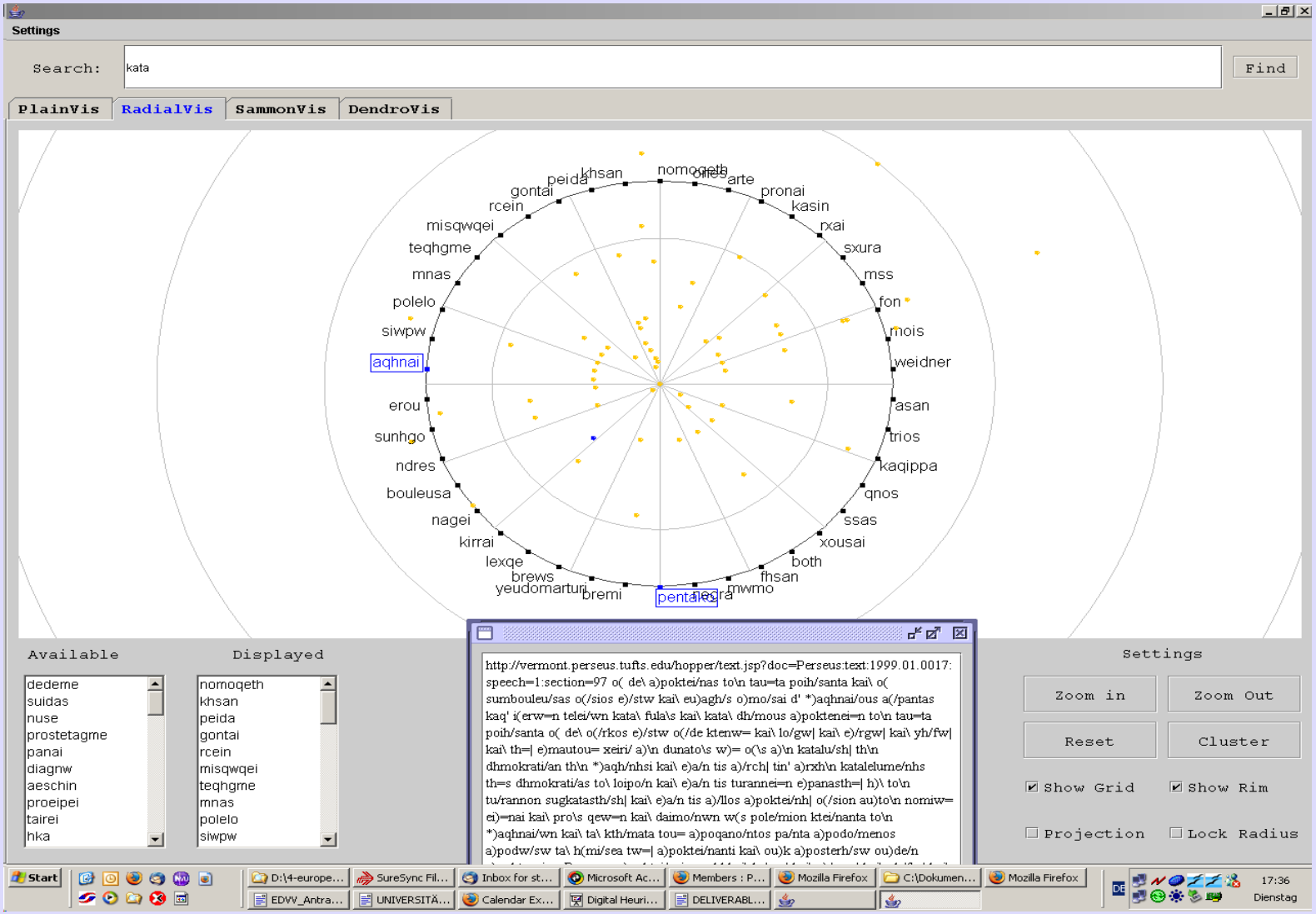
λαγών όνος *f.* [reltd. λαγαρός]
• how to translate Call.5.88 breasts and hips of Athena? or lagones more generally for the body, or the middle area of the body; perh. waist here. Sense 'womb', see Rev.Suppl., prob. doesn't exist.
• at Theoc.22.202 the spear pierces the lagwn and the omphalos: side/midriff and navel? Gow: the unprotected part of the abdomen between ribs and hips
• Plu.Arat.22: is lagwn here a hollow, recess, or just the side of a mountain? Are there really two senses for mountains? sense 2 seems to be needed for the cup and quiver.

λαγῷος ον *adj.* [λαγώς]
• does uncontr. lagwios exist neut.pl.sb., Ar.V.709 -- delicacies as LSJ, or every kind of hare dish, plenty of hare meat

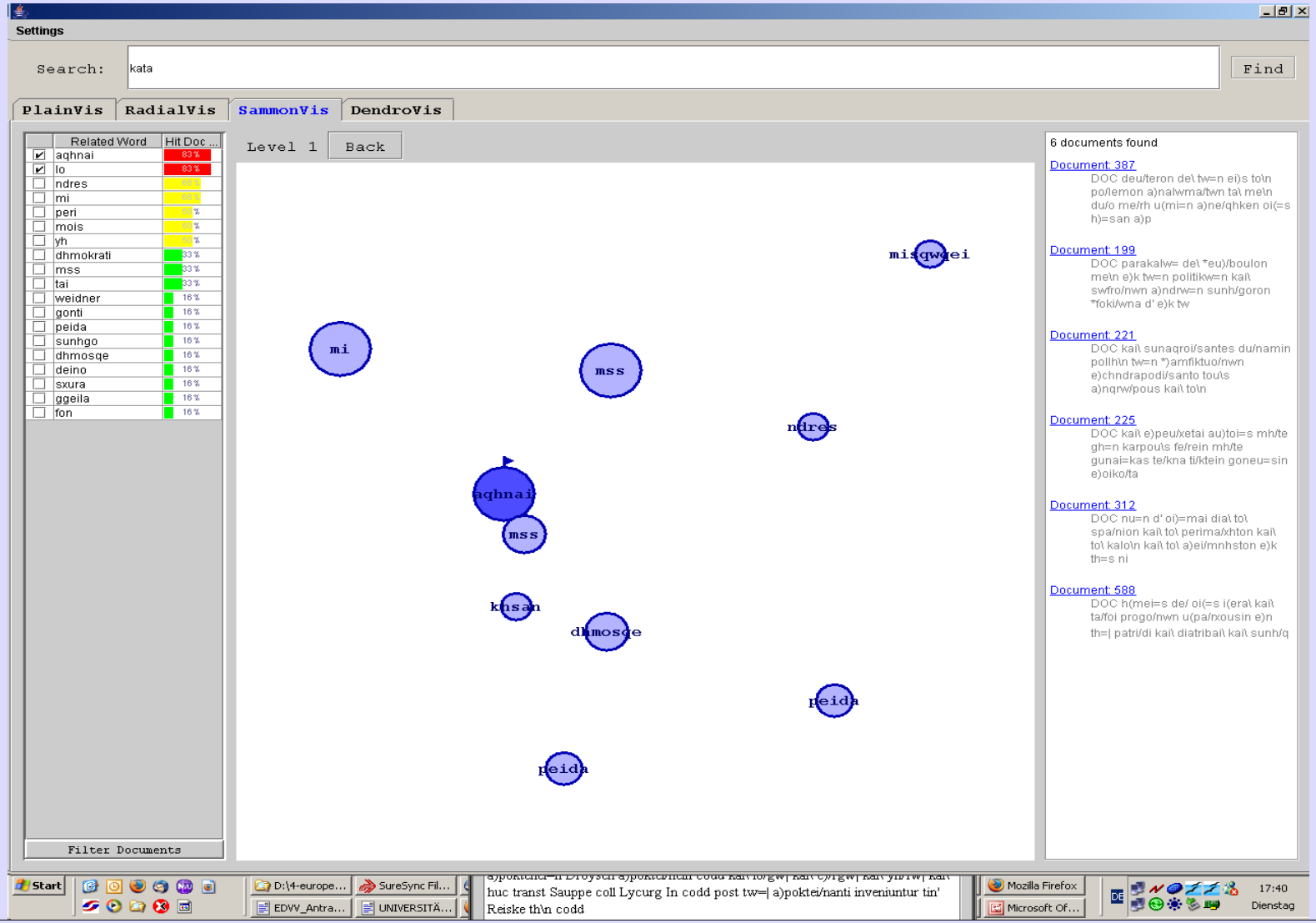
λαγώς λαγώ,, also **λαγῶς** λαγῶ, ep. **λαγωός** οῦ, Ion. **λαγός** οῦ *m. and f.* | acc.sg. λαγών (Ar.) |
• does this include rabbit? kuniklos is late. (mod. kouneli; lakoudaki is bunny i.e.rabbit, not small hare?)

(quickly?) and put it on. i.e. intensive of a difft. sense of lamabanw from grip a tool etc.
• it's difficult to assess how many exx. have the sense of 'pick up, take' and how many are just 'hold'.
• perh. sense 2 is wrong, it just means 'take, accept, pick up' and there is no intensive (eagerly or sim.). The garment ex. (Theoc.15.21) could then go here, pick up the garment (in order to put it on to go out) 1 and 2 could perhaps all be combined as one section
• Ar.Lys.209 sexual double-entendre here? most of the uses of this vb. would suit this. But in which section shd. this passage be placed? w.gen., construction needs to be pointed out -- but sense: each woman holds on to one part of the kulix, or each holds it in turn, or each drinks some of the wine in turn??
• There seems to be a connection betw. this ctxt. and Theoc.18.46: we will first draw fr. the silver flask and let drip smooth oil beneath that shady plane. (Gow.) Do they take the ointment \ or oil from the olpis into their hands, and then smear / or pour it in drops ? Or does the taking just loosely ref. to picking up the olpis along with the aleiphar.
• Does this passage lead us to think the Ar. passage

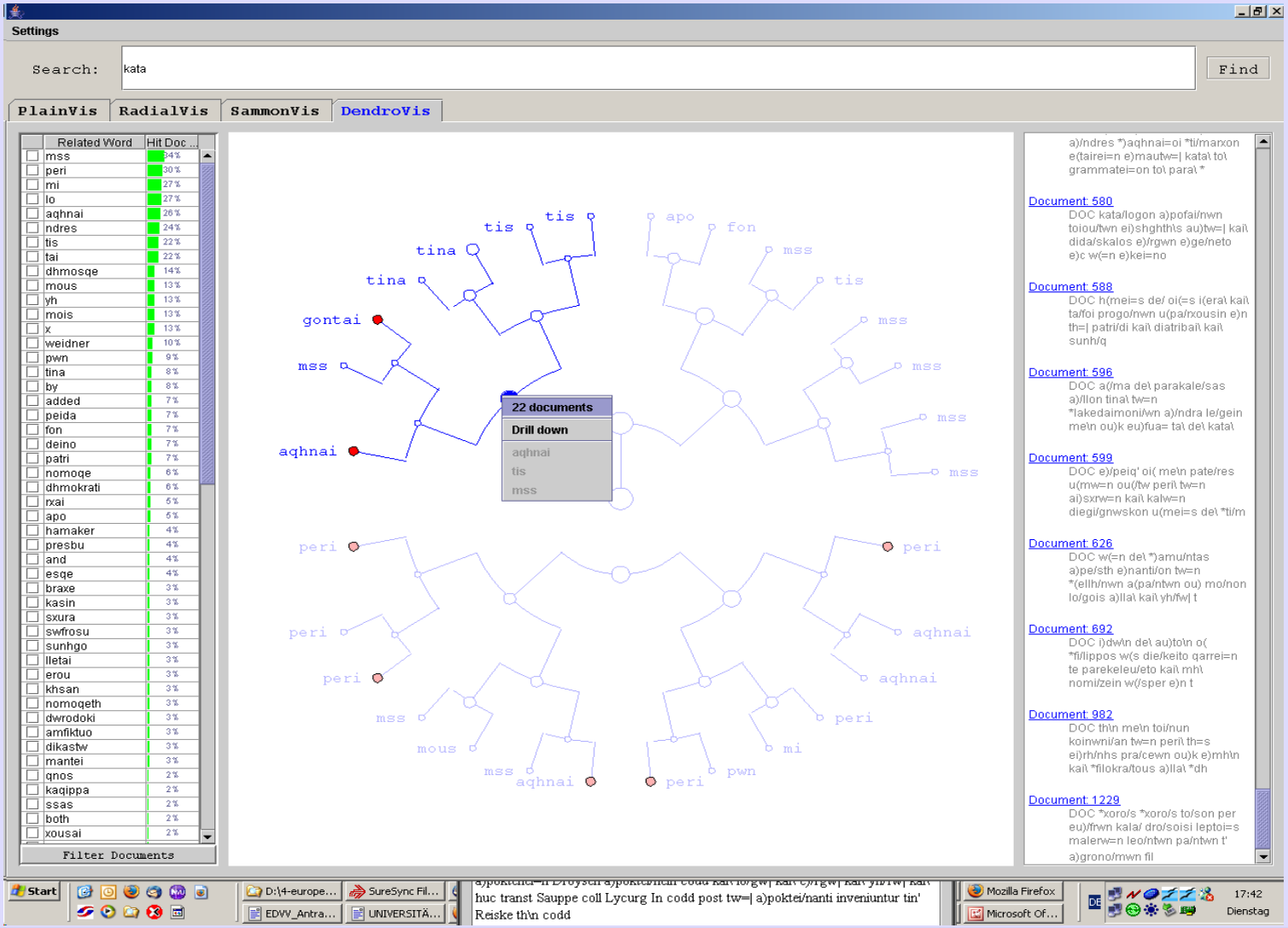
Lexicon <=> Corpus Visualisation I



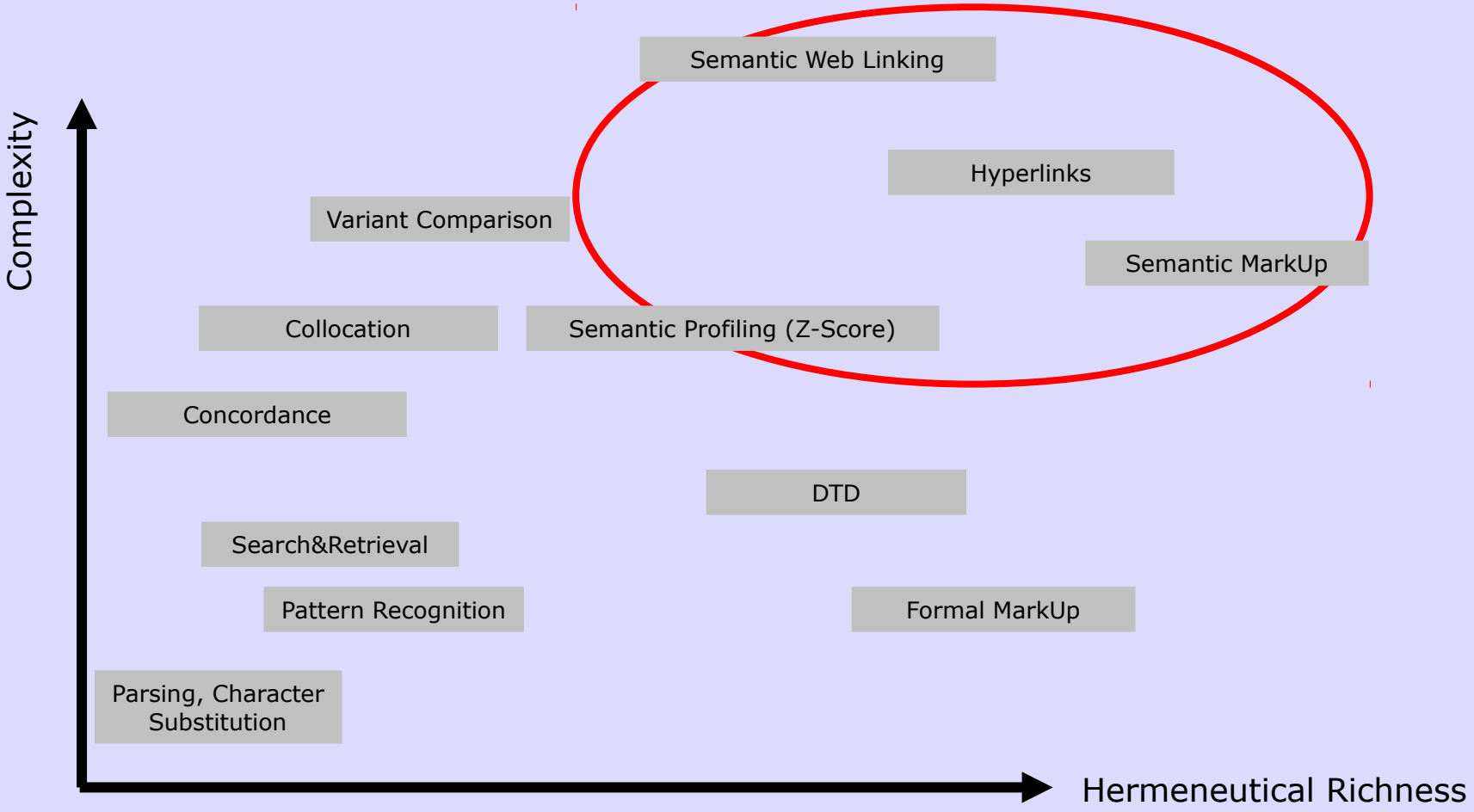
Lexicon <=> Corpus Visualisation II



Lexicon <=> Corpus Visualisation III



Digital Document Value Add-On
(c) J.-C. Meister



Conclusions I: Le document retrouvé ... Open Documents and Open Source in eScholarship

- Re-constructing the 'document' entity in networked, digital settings is a constitutive effort for eScholarship to work at all. It will need to combine
  - complex object modeling methodology (e. g. ORE, Europeana)
  - semiological models for complex information entities (RTP-DOC)
  - a deeper understanding of digital semantic networks (are these 'languages'??)
- For eScholarship to work at all a very specific understanding of the term 'open source' needs to be consequently and systematically applied: free availability of all source material!
  - Hence the primary characteristic of cyberinfrastructure as seen by the ACLS: *"It will be accessible as a public good"*
- The heuristics used for corpus modelling and aggregation as well as their technical implementations and foundations need to be open source in the more traditional sense of the term as well as based on open standards!
  - Hence ACLS recommendation 7: "*Develop and maintain open standards and robust tools*"

Conclusions II: OA/OS in eScholarship, Data vs. Publication

- The mainstream of the e-science OA discussion almost completely bypasses e-scholarship in that issues of publication economy and ease of access to journal articles are of minor relevance in our sector
- Instead, the need for OS and OA in e-scholarship stems from the needs of the rapidly evolving paradigm for digital work on source corpora in digital scholarship: OS and OA are key enablers for this new paradigm!
  - Hence recommendation 2 of the ACLS report: *"Develop public and institutional policies that foster openness and access."*
- In such a perspective, the technical separation of published results and of raw data (= source material) makes less and less sense: published results cannot be apprehended without the source material being available!
- Source data and publication formats tend to be even more entangled and considering them separately is not very useful: rather think about them as **one information continuum** with **several aggregation and abstraction layers**!

Conclusions III: Digital Representation vs. Language. 'Two Cultures' revisited (with some help from J.C. Meister, again)

? - string_chars(`[77,97,110,32,107,111,101,110,110,116,101,32,100,10
5,101,32,77,101,110,115,99,104,101,110,32,105,110,3
2,122,119,101,105,32,75,108,97,115,115,101,110,32,9

X=[One could divide humans into two classes and distinguish those understanding a metaphor from those understanding a formula. Those understanding both are too few to constitute a class of their own.]

(Heinrich von Kleist)

0,101,115,32,118,101,114,115,116,101,104,101,110,44
,32,115,105,110,100,32,122,117,32,119,101,110,105,1
03,101,59,32,115,105,101,32,109,97,99,104,101,110,3
2,107,101,105,110,101,32,75,108,97,115,115,101,32,9
7,117,115,46] `,X).

- The challenge is to **falsify Kleist**: in the end, much of the above may not be as specific for eScholarship, the two cultures divide may ultimately turn out artificial and we may well learn from each other, also regarding OA/OS.
- In this sense, the least we share is Rolf-Dieter Heuers concluding remark "Exciting times are ahead!"